14.310x Flipped Classroom materials

——all w3 —-

rduranl

Course Guide

14.310x Flipped Classroom

Week 1 Instructions

Some title

Checklist

- ☐ Complete the INTRO TO R interactive course from the `Swirl` package[1] (Requirement)
- ☐ Watch the Getting started with Google Colab notebooks video tutorial (Requirement)
- ☐ Create or set up a personal Google account
  (you must be able to use Drive and Colab). (Requirement)
- ☐ Create your first Colab notebook (Session 1)
- ☐ Complete Coding Lab 1. (Session 1)
- ☐ Complete Coding Lab 2 (Session 2)

C.1.0

Your first Colab notebook

Instructions: Read and follow the steps below **before proceeding with the activity**. After reading the instructions, access the notebook link and complete the exercises in Colab. This is an individual task, but you will collaborate on the final question.

Notebook: Your first Colab notebook

[1] **Module 1** > **Introduction to R** in the online component of the course.

We will cover the essentials of working with Jupyter Notebooks on Google Colab — this resource will be an important tool throughout. Before you can begin working on the coding labs in Colab, make sure to:

- **Create or use a personal Google account**. While it is possible for you to download the Jupyter Notebooks we will work on, and manage them locally in your own machine, we strongly advise you to work on Colab since it will make it easier for your classmates and the instructor to interact with your work when needed. Colab has many features Google Drive already offers: comments, real-time colaboration in the same document or folder, etc.
- **Save the shared notebooks and documents to your own drive.** The notebooks you will have access to are read-only; in order to work on them you will have to copy them to your drive. It is recommended that you maintain a per-week structure in your folders as this will make it easier to follow instructions, especially when reading files or collaborating with others. To save the notebook to your Drive, go to the File menu: `File > Save a copy in Drive`.
- **(When creating a new notebook) Change Colab Notebook's runtime type to R.** By default, when you create a new notebook in Colab, the virtual machine Colab sets up for you is a `Python` installation. This means all the cells in the notebook will only recognize and run `Python` syntax or commands. To switch to an `R` language setup:
  1. Open the notebook menu and go to `Runtime > Change runtime type`.
     - Or click the toggle in the upper-right corner (as shown below) and select `Change runtime type`
  2. In the dropdown labeled `Runtime type`, select `R`.
  3. Click `Save`.

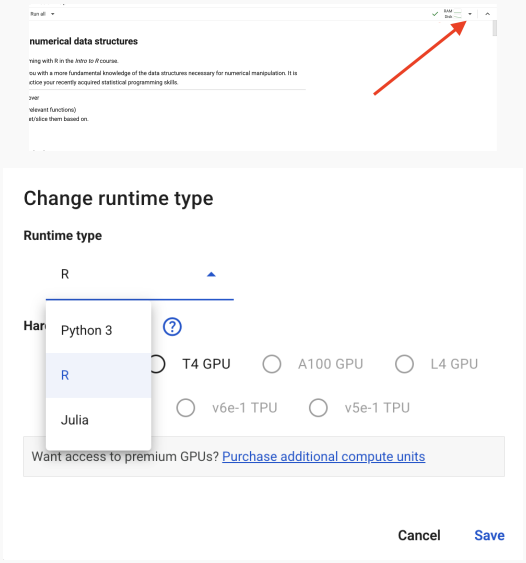


Fig.Changing the runtime type

You may now create your first Colab notebook

C.1.1

Coding Lab 1— Numeric data structures in R

Instructions: Work individually. Solve all exercises in the corresponding Colab notebook. Record your answers and/or code in your copy of the notebook.

Notebook: Coding Lab 1

Submit your work in the format required by the instructor.

C.1.2

Coding Lab 2 — Data manipulation with `dplyr`

Instructions: Work individually. Solve all exercises (sections 0 to 3) in the corresponding Colab notebook; record your answers and/or code in your own copy.

Notebook: Coding Lab 2

Optional **Section 4** allows work in pairs or teams of 3.

Upon completion, submit your work in the format required by the instructor.

14.310x Flipped Classroom
Week 2 Instructions

Some title

Checklist

- ☐ Complete the ADVANCED R interactive course from the `Swirl` package and watch the `ggplot` tutorial[2] (Requirement)
- ☐ Watch the **Import Data** tutorial[3] (Requirement)
- ☐ Read the note on importing `aiwars` and other datasets (Requirement)
- ☐ Complete Coding Lab 3 (Session 1)
- ☐ Complete Guided Case 2.3 (Session 2)

Importing `aiwars` and other datasets

Depending on flipped classroom logistics of your group, your access to the sessions' assets (datasets, figures, scripts, etc.) will come in one of two forms:

- **Through a direct URL** for instance, the original `aiwars.csv` dataset lives directly in this URL:
  `https://docs.google.com/spreadsheets/d/1NeZZWI2fT71M9QD8zjnz817T1B3XBM6lolqArSe9CwQ/export?format=csv`
- **Your instructor will provide them to you privately** and they will either:
  - Share them via a URL similar to the one above.
  - Share the files through other means.

[2] **Module 2 > R Course and R Tutorial: ggplot** in the online component of the course.
[3] **Module 3 > R Tutorials: Basic Functions** in the online component of the course.

As they may want to slightly modify the original files for grading purposes, or have any other goal in mind.

If a dataset's URL is provided, you can read the data directly into R with the URL and the appropriate reading function; simply provide the URL as the path. For instance, `aiwars` is in `.csv` format:

```
URL_aiwars <- "https://docs.google.com/spreadsheets/d...."

aiwars <- read.csv(URL_aiwars)
```

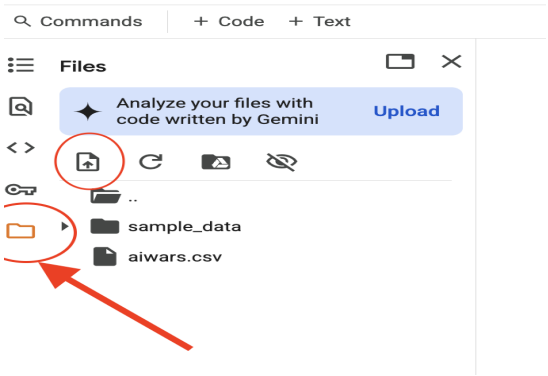
And similarly for other formats. Suppose we had a single Excel sheet:

```
install.packages("readxl")
library(readxl)

URL_aiwars_xl <- "https://...some-URL"

aiwars <- read_excel(URL_aiwars_xl)
```

If the file is shared in any other way, we will first need to upload it to the virtual machine's disk in Colab and read it from there, providing the `path` to it — as you would do if you were reading data to R in your own computer. To upload a file in Colab, click the folder icon in the leftmost menu bar, as shown in the screenshot. Then click the upload icon (circled in red); as you can see, we already uploaded the file.



Once uploaded the file can be read as a if in your machine:

```
aiwars <- read.csv("aiwars.csv")
```

Important: Colab runtime limitations

While your Colab notebook — including all code cells, text, and outputs — will remain saved, the underlying *runtime* (i.e., the virtual machine that executes your code) is temporary. After a period of inactivity or when a time limit is reached, the runtime will automatically disconnect.

When this happens:

- All variables and objects stored in memory (e.g., your R data frames, models, vectors) will be lost.
- Any files you uploaded manually will be erased.

When reconnecting, a fresh virtual machine will be started, and you'll have to re-run your code, and re-upload any needed files. Disconnects may happen occasionally as you work in off-notebook tasks; re-reading or re-uploading files should not take over 30 seconds.

Coding exercises in case studies

Throughout the course, the case studies you will work on contain a mix of conceptual and coding questions. To simplify your workflow, all the questions where you're required to write code are specially labeled. They appear in this format:

Question 5. [2.3] calculate the probability of event A...

The small icon signals that the question requires coding. The number inside the brackets (2.3 in this example) is a numbering **within the Session's dedicated Colab notebook**, which will often be different from the overall question number for the activity.

Further, if you click on the orange icon, it will take you directly to the question's cell within the Notebook by opening a new browser tab – remember you must work on your own copy of the notebook. These links are simply provided as a convenience to help you locate and visualize the relevant task quickly.

Apart from being the place where you are expected to code your answers, the corresponding cell usually includes a placeholder for your code, often accompanied by additional hints, extended context, or partially completed code to help you get started. When working on a case study, keep both this PDF and Colab notebook open and use these references to move smoothly between the written materials and your code.

C.2.3

Coding Lab 1 — Web Scraping in R: step by step guide

Instructions: Work individually. Solve all exercises in the corresponding Colab notebook. Record your answers and/or code in your copy of the notebook, or in the format required by the instructor.

Notebook: Coding Lab 1

Case study G.2.1

Random variables in the wild: *Reddit* posts and empirical distributions

Instructions: Work in pairs or groups of three. Answer the questions

Notebook: Case Study 1

Dataset: aiwars.csv

One of you must share editor access to their notebook with the rest.

Scenario: We will cover this week's lecture contents using a real-world dataset from Reddit. You will put your data manipulation (with dplyr and base R) skills to practice

---

Context: The `AIwars` dataset consists of a collection of posts scraped from the `r/AIwars` subreddit, a forum where users debate the societal implications of AI. These range from predictions about AI-driven job loss and technological conflict to satire, trolling, and speculation. The dataset captures a rich period of discussion and polarization. Each observation corresponds to a single post, with information about its author, contents and engagement.

Key Variables:

- `post_index` — Unique identifier for each post (starts at 1 and consecutively numerates all posts).
- `author` — Reddit username of the post author (may be `[deleted]`).
- `post_date` — an R `Date`
- `fulltext` — Full text of the Reddit post in format `TITLE: [some post title] TEXT: [some post text]`
- `post_length` — the net number of characters in `fulltext` (excluding the TITLE and TEXT headers).
- `post_upvotes` — Number of upvotes (user endorsements or *likes*) the post received.
- `comments_number` — Total comments the post received (includes replies to other comments).

Part 1. Basic dataset facts

1.1 [1.1] Install and load the `tidyverse` packages. Create `aiwars_URL`, a `character` with the download URL for `aiwars.csv`. Then read in the dataset as `aiwars`, a data frame. Use `glimpse()` to answer:

(a) How many posts are we working with?
(b) How many variables does the dataset have?
(c) How many are *true* `character` types?
(d) How many are numeric?
(e) Which characters are better described as factors? Why?

1.2 [1.1.1] For the entire case study, will use only the **Key Variables** described above. While `select()`ing columns from the current data frame is possible, a memory-efficient alternative is to read in only those columns we need.
Use the `col_select` argument in `read_csv()` to create `aiwars_short`, a data frame containing only the columns described as **Key Variables**.

1.3 [1.1.2] Use `select()` to create `aiwars_short2`, a dataframe with only the **key variables**.

1.4 Moving forward we will work with `aiwars_short` only. You may delete `rm(aiwars, aiwars_short2)`.

1.5 Use `summary()` or other summarizing methods to answer the following questions:

(a) What is the time span of the posts?
(b) On average, how long (in words) is a post on this subreddit?
(c) How many posts don't have any replies?
(d) What is the largest number of upvotes in a post?
(e) Which author has posted the most? (Consider `as.factor()`).
(f) How many distinct users have posted in this subreddit? (Consider `unique()`).

Part 2. Counting posts

2.1 [2.1] Create the following variables in `aiwars_short`:

(a) `popular`, takes value 1 when the post has 29 or more upvotes. Takes value 0 when the post has less than 29 likes.
(b) `text_classification`, which takes value `mostly title` if it is less than 70 characters long, `short` if it has 70 or more but less than 110 characters, `common` if it has 110 or more but less than 900 characters, and `long` if the post has over 900 characters.

2.2 How many `common` posts are there?

2.3 If we draw a post at random, it will be a popular one with what probability?

2.4 If we instead sample 10 posts at random **with replacement**, what is the probability 3 of them are popular?

2.5 [2.2] Suppose random variable X; $\Omega_X = \{0, 1, 2, ..., N\}$ represents the number of popular posts in $N = 10$ draws with replacement from `aiwars_short`. For the following probability statements describing various events, first write them down as a probability and then use R to compute them.

(a) What is the probability of getting at most 4 popular posts?
(b) What is the probabililty of getting at least one but at most 3?
(c) What is the probability of getting a number of popular posts that is not 0 or 4?

2.6 Similarly, we can treat the texts' classification as random variable Y. Is it continuous or discrete? What is Ω_Y?

2.7 [2.3] Express $f_Y(y)$ as a data frame `f_Y` with as many rows as elements in

Ω_Y and two columns: `y`, the value taken by Y and `p` its mass or density.

2.8 What is the probability a post is not `mostly title` ?

2.9 What is the probability a post is either `common` or `long`?

Are popular posts different?

2.10 What is the probability a post is popular **AND** `long`?

As we learned from the lecture, if two events are **not independent**, the additional information one of them provides should *update* the probabiility of the other. We will examine this fact by calculating the probability of a post being long **GIVEN that we know it is popular**:

$$P(Y = long \mid popular = 1)$$

As a reminder:

$$P(long \mid popular = 1) = \frac{P(long \cap popular = 1)}{P(popular = 1)}$$

2.11 Use your answers to questions 2.3 and 2.10 to compute the probability a post is long GIVEN it is popular. How does this probability compare to the unconditional probabiity $P(Y = long)$?

2.12 We now know that the event $(Y = long)$ has two mutually exclusive and exhaustive partitions: $(popular = 1)$ and $(popular = 0)$. Following the lecture, Show that the Law of Total Probability verifies for this event.

2.13 What is $P(popular = 1 \mid long)$? Is your answer the same as in 2.11 ? Should it be? Why?

2.14 [2.4] Create `joint_pmf`, a data frame similar to the one you created for 2.7; in this case, it should display the probability for all possible values in $Y \cap Popular$. This is the ordered pair $\Omega_Y \times \Omega_{Popular} = (long, 1), (long, 0), ..., (\text{mostly title}, 1), (\text{mostly title}$

2.15 [2.4.1] Use `joint_pmf`, `p` (or $f_{popular}$) and `f_Y` to compute the following in R:

(a) $P(Y \mid Popular = 1)$, a vector of length 4.

(b) $P(Popular \mid Y = common)$. What is its length? .

14.310x Flipped Classroom

Week 3 Instructions

Some title

Checklist

- ☐ Complete Guided Case Study 2 (Sessions 1 & 2)
  - ☐ We will continue to work with `aiwars`
  - ☐ Get the `aiwars_embeddings` dataset here
- ☐ Complete Open Case Study 1 (Session 2)

Case study G.3.2

Analyzing text similarity:
language semantics and random variables

Instructions: Work on your own for Session 1, and in pairs or groups of 3 for Session 2.

Notebook: Case Study 2

Dataset: `aiwars_embeddings.csv`

Part 1 Part 2 Part 3

Dataset: `speech_anchors.csv`

The `aiwars_embeddings` dataset (split into three parts for file-sharing purposes) contains 1024 variables associated with each post's `fulltext`. You will learn what these embedding variables `V1`, `V2`, ..., `V1024` represent below, but keep in mind `post_index` in both datasets uniquely links each post in `aiwars` to its embedding in `aiwars_embeddings`. Dataset `speech_anchors` is a 2-row data frame with the same structure.

Part 0. Required data and packages

0.1 [0.1] Load/install the `tidyverse` and the `geometry` packages.

0.2 [0.1] Read in `aiwars`, the same dataset as last week. Also, read in `speech_anchors.csv` as `anchors`. This should be a data frame with two rows and 1025 columns.

0.3 [0.1] Read in parts 1,2 and 3 of `aiwars_embeddings` as `part1`, `part2`,`part3` respectively. Then use `rbind()` to vertically merge the parts into `aiwars_embeddings`. Optionally, use `arrange()` to sort the dataset in `post_index`- ascending order.

Scenario:

Natural Language Processing (NLP) focuses on enabling machines to understand and work with human language. One of the key advances in NLP has been the development of methods that represent text — such as words, sentences, or entire posts — as vectors. These vector representations, often referred to as *embeddings*, capture aspects of a text's semantic content: its meaning, tone, and topical focus.

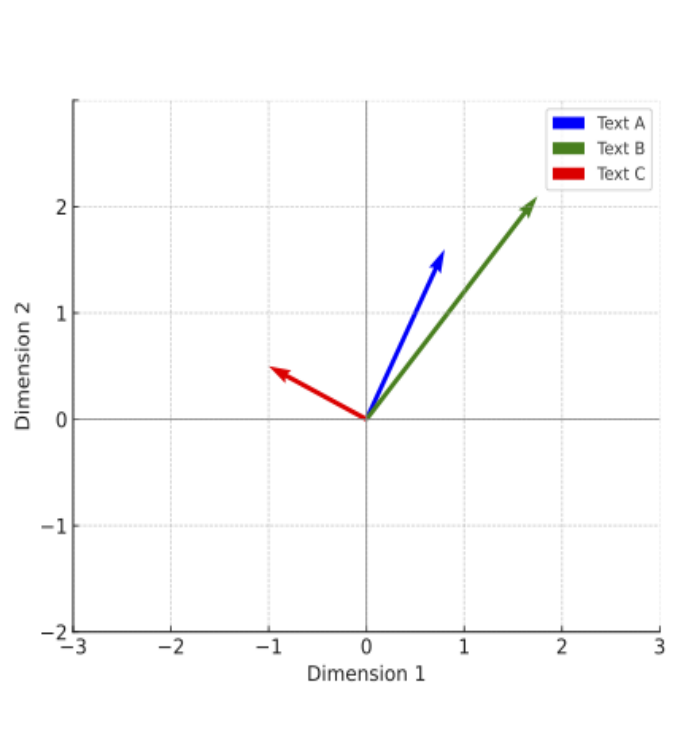
To illustrate the idea, suppose we have the following texts:

- Text A: "I love how these new AI tools can generate art!"
- Text B: "AI-generated images are cool, but I'm worried about copyright."
- Text C: "I painted this myself — no software involved."

We embed these texts in 2-dimensional vectors from the origin, plotted below:

- $\vec{a} = \begin{pmatrix} x_a \\ y_a \end{pmatrix} = \begin{pmatrix} 0.8 \\ 1.6 \end{pmatrix}$
- $\vec{b} = \begin{pmatrix} x_b \\ y_b \end{pmatrix} = \begin{pmatrix} 1.7 \\ 2 \end{pmatrix}$
- $\vec{c} = \begin{pmatrix} -1 \\ 0.3 \end{pmatrix}$

Figure 1: **2D text embeddings (vector from the origin)**



Despite their differences in tone, $\vec{a}$ and $\vec{b}$ in Figure 1 are close to each other because the texts they represent use AI-related language, whereas $\vec{c}$ is far removed given the lexical differences. In part 2, we will explain what "close to each other" means for vectors, with two common semantic similarity measures: **direction** (i.e., the *angle* they form) and **distance** between them.

Part 1. Coding vector operations

You will start by programming functions necessary for this analysis, such as the *magnitude* of a vector, the *distance* between two vectors, their *dot product*, etc.

Distance between two vectors

The distance from $\vec{a}$ to $\vec{b}$ in Figure 1 is defined as:

$$|| \vec{a}, \vec{b} || = \sqrt{(x_a - x_b)^2 + (y_a - y_b)^2}$$

This is the sum of their x and y-coordinates' differences squared (in any order). Finally, we take the square root of that sum. Substituting the values in the scenario description:

$$\cdots = \sqrt{(0.8 - 1.7)^2 + (1.6 - 2)^2}$$

$$\cdots = \sqrt{0.81 + 0.16} = 0.985$$

1.1 [1.1] Save $\vec{a}, \vec{b}, \vec{c}$ as numeric vectors `a`, `b` and `c` respectively. Compute the distances for each pair of vectors and save them accordingly: `dist_ab`, `dist_ac`, `dist_bc`. Hint: you can apply operations on all the coordinates at once: `(a-b)`.

(a) Make sure your result for `dist_ab` is the same as above.

In general, to get the distance between any two n-dimensional vectors

$$\vec{V} = \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{pmatrix}, \vec{U} = \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix}$$

we generalize the process with all their coordinates and take the square root:

$$|| \vec{V}, \vec{U} || = \sqrt{(v_1 - u_1)^2 + (v_2 - u_2)^2 + \cdots + (v_{n-1} - u_{n-1})^2 + (v_n - u_n)^2}$$

$$\cdots = \sqrt{\sum_{i=1}^{n} (v_i - u_i)^2}$$

1.2 [1.2] `aiwars_embeddings` contains a 1024-dimensional vector per post. Program a distance function `dist_function(V, U)` that takes any two equal-length, numerical vectors. It returns the distance between them. If you didn't in the previous exercise, make sure to use vectorized operations (e.g., `sum()`, squaring all elements at once as $(a-b)^2$) to generalize the computation.

(a) Verify your function is correct and reproduce the results in 1.1

Distance from one, to multiple vectors

Working with individual, separate objects such as `dist_ab`, `dist_ac`, and so on can be difficult and tedious. In R, we will often want all the distances with respect to $\vec{a}$ in a single "distances" vector or variable:

$$D_a = \begin{pmatrix} || \vec{a}, \vec{a} || \\ || \vec{a}, \vec{b} || \\ || \vec{a}, \vec{c} || \end{pmatrix}$$

We will now pass `dist_function()` trhough all the example vectors at a time to obtain D_a — the steps should be familiar from the previous Coding Lab.

1.3 [1.2.1] Create vector `D_a` with a *for* loop by following the steps below:

- Declare `D_a`, an empty vector.
- Use `rbind` to create `examples_matrix`, a matrix with 3 rows and 2 columns. One row for each example vector `a,b,c` and one column for each of their x, y-coordinates.

- Declare a `for(...)...` loop with as many iterations as rows in `examples_matrix`.
- Inside the *for* loop, code the following:
  - Subset `examples_matrix` to obtain vector `v`.
  - Obtain the `distance` between `a` and `v` with `dist_function`.
  - Use `append` to recursively add a `distance` on each iteration to vector `D_a`.

1.4 Verify your answers with those of 1.2. What should be the result for $||\vec{a}, \vec{v}||$when $\vec{v} = \vec{a}$? Why?

Finally, we can wrap the code we created for 1.3 in a function that directly returns D_a when given a vector and a matrix.

1.5 [1.2.2] Create a function `get_distances(u, M)` which takes a vector `u` of length n and a matrix `M` with an arbitrary number of rows k, and n columns. The function should `return` D_u, a vector of length k where each element is the distance between `vector` and a row of `matrix`.

1.6 Use `get_distances(_,_)` to reproduce 1.3. Verify it is the same vector.

The norm of a vector

The norm $|| \vec{V} ||$ (also called the *magnitude* or *length*) of a vector indicates how long the arrow is from the origin to the point it reaches. Since all embeddings are vectors at the origin (all their starting coordinates are 0), we can calculate this as the distance between a vector $\vec{V}$ and the n-dimensional zero: $||\vec{V}, \vec{0}||$.

$$|| \vec{V}, \vec{0} ||=|| \vec{V} ||= \sqrt{(v_1 - 0)^2 + \cdots + (v_n - 0)^2}$$

$$... = \sqrt{\sum_{i=1}^{n} v_i^2}$$

1.7 [1.3] Create `norm0_function`, a function that takes `V`, a numerical vector of length n as an input, and returns its norm from the origin. Hint: you can recycle `dist_function` inside `return()` with the appropriate `U`.

1.8 [1.3.1] Compute $|| \vec{a} ||, || \vec{b} ||$ and $|| \vec{c} ||$. Save them to `a_norm`, `b_norm`, `c_norm` respectively.

1.9 Similar to 1.3 and 1.5, create a function `get_norms0(M)` taking a matrix of n rows/vectors, all with the same number of dimensions/columns m, and returns a vector N_0 of length n with a norm from the origin for each row/vector in `M`. Input it the `examples_matrix` and verify your results in 1.8.

The dot product of two vectors

The dot or *scalar* product of two vectors $\vec{U} \cdot \vec{V}$ is defined as the sum of the product of each pair of their coordinates:

$$\vec{U} \cdot \vec{V} = u_1 * v_1 + u_2 * v_2 + \cdots + u_n * v_n$$

$$= \sum_{i=1}^{n} u_i v_i$$

You can use R's native vectorized multiplication: `sum(u*v)`.

1.10 [1.4] Create a function `dot_product(u,v)` that takes two numeric vectors of the same length as input and returns their dot or scalar product.

1.11 [1.4.1] As in previous exercises, we're interested in computing a vector $\mathbb{D}_a$ of dot products. Use a *for*-loop and wrap it in a function `get_dot_products(v, M)` taking a vector `v` of length n and a matrix `M` of as many rows as vectors and n columns and returns a dot product $\vec{V} \cdot \vec{m}$ for each row.

Part 2. Processing AIwars posts' text embeddings

We prepared the `aiwars_embeddings` dataset for you by using one of OpenAI's text embedding products[1]. For each post's `fulltext` in `aiwars`, there is a vector representation of length $n = 1024$ in `aiwars_embeddings`:

Variables

- `post_index`: a consecutive index **in the same order as** `aiwars`. A given post and its embedding share the same post index.
- `V1`, `V2`, ..., `V1024`: the coordinates $v_1, v_2, ..., v_{1024}$ of each vector embedding $\vec{V}$.

You will now apply the functions you previously programmed to measure the semantic similarity of real-world embeddings.

Semantic similarity measures: *Luddites* vs *Synthists*

To illustrate semantic similarity measures, we will take the embeddings from four select posts from the AIwars subreddit, each with distinct tones and topics:

2.1 [2.1] Filter `aiwars` for posts with the following `post_index` values: 456, 584, 2397, 2526. Select the `fulltext` and post index columns only. Assign the resulting dataframe to `tones`.

2.2 Skim the title and body of each of the posts in `tones`.

2.3 Create a character vector `post_labels <- c("luddites", "review", "critique", "synthists")`. Sort `tones` in ascending `post_index` order and add the vector as a new variable `label`.

2.4 [2.2] Retrieve the embeddings for `tones` in a 4×1024 matrix and save it as `tones_embeddings`:

(a) Filter `aiwars_embeddings` for the relevant `post_index`.

(b) Either drop `post_index` or tidy-`select` all variables `starts_with()` "V".

- Remember to keep track of the embeddings. We suggest you previously arrange by ascending post index.

(c) Coerce the resulting data frame into a matrix with `as.matrix()`

.

Distance-based semantic similarity

As exposed in the introduction, if two texts share style, vocabulary or topics, their embeddings will tend to be closer, and on the contrary, distinct semantics will produce embeddings that are farther apart.

2.5 [2.3]

Table 1: Pairwise distances between example embeddings (not the real values)

	luddites	review	critique	synthists
luddites	0.0000000	72.384921	18.532947	91.004327
review	72.384921	0.0000000	56.110239	8.231476
critique	18.532947	56.110239	0.0000000	44.873291
synthists	91.004327	8.231476	44.873291	0.0000000

Code a 2-D distance matrix like the one displayed above with the embeddings in `tones`. Hints:

- This is a matrix with `rownames` and `colnames`
- Use `rbind` to save each iteration on `tones`.
- Note: the matrix you get should be symmetrical. The distance between *critique* and *luddites* should be the same in (row 1 column 3) and (row 3 column 1).

2.6 Before you reflect on the results, consider what the bounds could be:

(a) If the embeddings could be of any magnitude, what would be the maximum distance between embeddings?

(b) Compute the norm about the origin for the embeddings in `tones`. What can you notice?

(c) Compute the norm for all of `aiwars_embeddings` by getting a matrix for all the embeddings as in 2.4.

(d) Based on this observation, what is the longest possible distance between two of these embeddings?

2.7 Which embedding pair is the furthest apart?

2.8 Which individual embedding seems to be further away from the rest?

2.9 Which two embeddings are the closest?

Direction-based measures: Cosine Similarity

An alternative approach is to compare the vectors' direction only, through the angle θ formed by two vectors $\vec{a}, \vec{b}$. The cosine of said angle θ is defined as:

$$cos(\theta) = \frac{\vec{a} \cdot \vec{b}}{|| \vec{a} || \ || \vec{b} ||}$$

- $\mathbf{cos}(\theta) = \mathbf{1}$ means the angle is $\theta = 0°$, indicating identical direction. These texts will be very close in meaning and topics.
- $\mathbf{cos}(\theta) = \mathbf{0}$, the angle is $\theta = 90°$, indicating the posts are orthogonal, and for the most part, semantically unrelated: different topics and intent.
- $\mathbf{cos}(\theta) = \mathbf{-1}$, with angle $\theta = 180°$, vectors point in opposite directions.

In our setup, cosine similarities in $[0.5, 0.7]$ usually indicate a non-trivial topical relation; values ≥ 0.8 often reflect very close texts (near-paraphrases). Scores in $[-0.3,\ 0.3]$ are not indicative of a strong semantic link. ≤ -0.3 cosines can occur but are uncommon and shouldn't be interpreted as "opposite views" on their own—you should inspect the post instead.

2.10 [2.4] Write a function `batch_cossim(u,M)` taking a vector of length n and a matrix with k row-vectors and n columns, returning the cosine of the angle θ between u and each row-vector. Keep in mind what you learned in 2.6b and consider that class of vectors only.

2.11 [2.5] Similar to 2.5, create a cosine similarity 2-D matrix for the `tones` embeddings. The diagonal elements must all be equal to 1.

AI-awareness and Simplicity measured in the *aiwars* posts

For the next case study, we shift our attention to general qualities within the posts. We're interested in vector directions capturing meaningful, human–interpretable features. For the AIwars subreddit, a predominant topical direction is that of **AI awareness**. Stylistically, it is also possible to track a stylistic direction that contrasts short, slogan-like, low-context statements with longer, hedged, evidence-seeking discussion; we refer to this as *anti-nuance* or **simplicity**.

The embeddings corresponding to said features are contained in `speech_anchors`.

3.1 [3.1] Create vectors `ai_awareness` and `simplicity`, the cosine similarities between the `aiwars_embeddings` dataset and each of the `speech_anchors` previously discussed. Add them as variables of `aiwars` for your next case study — remember to previously arrange both datasets by ascending `post_index`.

3.2 save the new dataset as `aiwars_updated` to use during this week's next session.

Case study O. 3. 1

Instructions: Work in pairs or groups of three; solve the following exercises collaboratively, and put together the deliverable specified in Part 2.

Notebook: Case Study 2

Dataset: `aiwars_updated.csv`

One of you must share the case study's notebook and a blank slides or text document (part 2) with the rest.

---

[1] The coordinates for each embedding are outputs of a more complicated embedding model involving several neural networks architectures and outside the scope of this course. For this week's case studies will limit ourselves to work with the embedding themselves.

Part 1. Discrete and continuous methods

0.1 [0.1] Read in `aiwars_updated`; load/install the `tidyverse` packages.

Consider two continuous variables related to the semantics of text posted by redditors in r/Aiwars:

- **AI-awareness**: A, the extent to which a post's vocabulary and syntax depict artificial intelligence topics.
- **Simplicity**: S, a post's reliance on short, declarative phrasing and minimal argumentative or conceptual complexity.

1.1 Based on your previous knowledge, what are the sample spaces Ω_A, Ω_S ?

1.2 Provide a short explanation as to why these can't be discrete random variables.

1.3 [1.1] Use `ggplot2` to create A histogram plot for both A and S on the same x-axis.

(a) Declare a `ggplot()` on your data frame.

(b) Add a `geom_histogram()` layer with `ai_aware` as your x-`aes()`. Outside `aes()`, set the `fill` parameter to "blue", the `color` to "black", and the transparency (`alpha`) to 0.5.

(c) In the same plot, add another histogram layer analogously for `simplicity` but fill it with "red".

1.4 Examine the plot:

- What is the y-value?
- At what range of x-values does `ai_aware` have more observations?
- If we summed all y-values of the bins in the `simplicity` histogram, what should be the result?
- Can we tell exactly what $f_A(0.5)$ or $f_S(0.75)$ are? Why or why not?

1.5 More generally, which histogram seems more concentrated around a given range of values?

1.6 Are these histograms valid distributions?

1.7 We want the y-value to reflect a probability. Edit the `aes()` function in your histograms to include the following y-aesthetic: `y = after_stat(count/n)`. Compute and save `n` separately.

1.8 Similar to 1.4, what should be the sum of all the bins' y-values of any histogram?

1.9 Very roughly (no computation needed), what is x such that $F_A(x) \geq 0.5$?

1.10 Again roughly, what do you think $F_S(0.1)$ approximates to?

1.11 Histograms are discrete representations of a RVs. Change the `binwidth` of **just the red histogram** to 0.01. What can you observe? Did the distribution's relative shape change? Why?

1.12 Examine the new plot. What is the default bin width for the blue histogram then?

Recall the *Summarizing and Describing Data* lecture. We can get a **continuous** representation of our RV's distribution through kernel density estimation. Base R has a readily available function for this: `density()` — feel free to test it in your own time.

We will use `ggplot2`'s version: `geom_density()` to plot the estimated pdfs $\widehat{f_A(\cdot)}$ and $\widehat{f_S(\cdot)}$ for all the x-values in the domain of our data.

1.13 [1.1] Copy and paste your last histogram plot. Delete the `y` argument in `aes()` in both histograms. Substitute `geom_histogram` with `geom_density` in both layers.

1.14 Exclusively in terms of shape and position relative to each other, how do these distribution representations compare to those in 1.3?

1.15 Now examine the y-axis. How does it compare to your plot in 1.3?

1.16 Roughly, what is $max\ \widehat{f_S(x)}$? What is the minimum for both?

1.17 Are these valid distributions? To help you answer this question, consider random variable $M \sim U[0, \frac{1}{4}]$, a uniform distribution from 0 to 0.25; write down its pdf $f_M(m)$?

1.18 Roughly from the plot, what is $P(0.125 \leq S \leq 0.25)$?

1.19 Go back to your second plot in 1.3 and compute $P(0.125 \leq S \leq 0.25)$.

Another way to compute the probability of an event for a continuous RV is through its CDF. Recall from the lecture that

$$P(a \leq X \leq b) = \int_a^b f_X(x)dx = F_X(b) - F_X(a)$$

Base R allows us to create a function for the empirical CDF of continuous variables with `ecdf()`. This function also has its `ggplot2` analog in `stat_ecdf()`, which you can use just as any other `geom`.

1.20 [1.3] Use `ecdf()` to create functions `F_A` and `F_S`, the empirical CDFs of A and S respectively. Make sure you have the correct object by quickly plotting the CDFs: `plot(F_A)`, `plot(F_S)`. Then answer the following questions:

(a) $F_S(0.25)$?
(b) $F_A(0.25)$?
(c) What is $s \in \Omega_S$ such that $P(S \leq s) = 0.8$?
(d) What is $a \in \Omega_A$ such that $1 - P(A \leq a) = 0.1$?
(e) Compute the *extremes* event probability for A: $P(S \leq 0.36 \cup S \geq 0.6)$
(f) Make a 95% inner interval for S. Find any $a, b \in \Omega_S$ such that $P(a \leq S \leq b) = 0.95$

1.21 [1.4] Use `stat_ecdf()` in `ggplot2` to jointly plot the CDFs for A and S. Make sure to `color` them blue and red, respectively.

1.22 Can you conclude there's first-order stochastic dominance of one of the random variables to the other?

1.23 In the lecture, height was the construct measured for both Bihar and the US. Are S and A measuring the same construct?

1.24 What, if any, could the implications of FOSD be in this case?

Part 2. Non-independence

2.1 Within the r/AIwars subreddit, conditional on learning posts tend to be long and argumentative, would their probability of being AI-related change?

2.2 [2.1] Let's begin by examining $f_{A,S}(a, s)$. Use `ggplot2`'s function `geom_bin_2d()` to plot a 2D-histogram with `ai_aware` in the `x`-axis and `simplicity` in the `y`-axis.

2.3 What does color represent in this plot? Note that the "minimal" unit colored is a rectangle tile.

2.4 Are the x and y-bandwitdths the same? What are they, approximately?

2.5 Similar to 1.8, modify the color to show a probability with `after_stat()`. This time change the `fill` aesthetic.

2.6 Generally describe the joint composition of the subreddit's posts in terms of their AI-awareness and their simplicity.

2.7 Do short, declarative post tend to discuss AI extensively?

2.8 Do relatively long and argumentative posts necessarily discuss AI deeply and technically?

2.9 What is the semantic region $\Omega_S \times \Omega_A$ most posts fall in?

2.10 **True** or **False**: if A and S were independent, we would have a perfectly squared 1×1 grid and all tiles would be of the same color, because $f_{A,S}(a, s) = f_A(a) \cdot f_B(b)$

14.310x Flipped Classroom

Week 4 Instructions

Some title

Checklist

- ☐ Complete Coding Lab 4 (Sessions 1 & 2).
- ☐ (Optional) Explore the **Review Notes** and further **Readings** available.

Coding lab L.4.4

Understanding Probability and Statistics in R: A Step-by-Step Guide

Instructions: Work individually. Answer all questions in sections 1-6. Sections 7 and 8 are optional; follow your instructor's directions.

Notebook: Coding Lab 4

Record your answers in your copy of the notebook, or in the format required by your instructor.

Additional resources

– **Drive:** Review Notes

- **Drive:** Readings

14.310x Flipped Classroom
Week 6 Instructions

Some title

14.310x Flipped Classroom

Week 7 Instructions

Some title

Checklist

- ☐ Set aside this Wikipedia article on testing two-proportions hypotheses, in case you need it during Session 2 (Requirement)
- ☐ Skim this Wikipedia article discussing the Local Average Treatment effect (LATE) , in advance of Session 2. (Requirement)
- ☐ Complete Guided Case Study 3 invidually (Session 1)
  - ☐ Retrieve the `students` dataset here
- ☐ Complete Open Case Study 2 in pairs or teams of 3. (Session 2)
  - ☐ You will continue to use `students`

Case study G.7.3

Evaluating AI-assisted learning on student outcomes

Instructions: Work on your own; read the scenario and answer the questions. Type your answers in the format required by the instructor.

To work on the `students` dataset you may use either a Colab notebook or your own installation of R.

Dataset: students.csv

Your code will not be evaluated, but keep your R script or notebook tidy, as you may need to review some of your answers during Session 2.

Scenario:

You are the Government of *Novaria*'s new Minister of Education. The Prime Minister has tasked you with evaluating a primary education policy recommendation: the rollout of AI-assisted learning for mathematics curricula in grades 5-8.

The proposed program, *Project Mentor*, involves deploying a large language model (similar to ChatGPT) named *AlgebrAI* specifically trained and fine-tuned for elementary math tutoring. AlgebrAI's interface is tailored to deliver interactive, one-on-one tutoring sessions to students. The AI mentor adapts to each student's skill level and provides problem-solving guidance, hints, and feedback designed to help the students master their grade's math curriculum.

Each participating school receives a number of tablets with AlgebrAI pre-installed, configured for offline-first use and automatically synced with central servers when internet is available. Students selected for treatment attend 20-minute tutoring sessions per day under the supervision of a facilitator.

The Prime Minister believes Project Mentor can boost test scores nationwide, but political opponents have raised concerns over cost and long-term efficacy. You are now in charge of evaluating the impact of the program in grades. Your team provides you with:

1. A 6th-grade math test designed to perfectly measure domain of the curriculum in a scale from 0 to 100.
2. A list of 1,000 students enrolled in 6th grade across Novaria, picked at random — part of the `students` dataset. This list contains only the following variables:
   - `unit`: a consecutive number assigned to the student.
   - `W_school`: Indicates whether the student's school is managed by the government ($W_{school} = 1$) or if it is privately managed ($W_{school} = 0$)

You have authority to apply the exam to any 6th grader in Novarnia, and you can implement the program (tablet usage and monitor time) in all government-managed schools, but to include any students attending a private school to the program you must first obtain authorization from their school board.

Exercises

Consider $T_i \in \{0, 1\}$ the treatment status of student $i = 1, 2, ..., 1000$ — $T_i = 1$ if treated $T_i = 0$ if not treated. Potential outcomes $y_i(T_i)$ in `students`, measured in test results (grades 0 to 100) are defined:

- `y0`: vector $Y(0)$, assume we can't observe it unless specified.
- `y1`: vector $Y(1)$, assume we can't observe it unless specified.

1.1 What is the value of $y_3(1)$? Describe its meaning —in terms of the potential outcomes framework.
1.2 What is the value of $y_5(0)$? Describe its meaning.
1.3 Compute $\bar{Y}(1)$. Describe its meaning.
1.4 Suppose $T = 0$. What is the value of y_{20}^{obs} and y_{40}^{miss} ? Briefly explain why.
1.5 Suppose $T_i = 1$ for all $i = 1, 2, 3, ..., 1000$. What is $\bar{Y}^{obs}$?
1.6 Imagine you can observe both potential outcomes.
  (a) What is the causal effect of Project Mentor in student 245?
  (b) What is the estimated Average Treatment Effect (ATE) of Project Mentor?
  (c) Does the estimated ATE support the Prime Minister's claims?

In practice, only Y^{obs} will be available after applying the exam. You will observe **one test score per student**, as well as the students' treatment status: either treated or untreated.

After careful consideration, your team assigned N_0 students to control and N_1 students to treatment out of the $N_0 + N_1 \equiv N = 1000$ sampled. The assignment criteria included logistics, the school-year timeline, operation costs and potential political opposition. Treatment was allocated amongsts students per the rule

$$T = W_{school}$$

2.1 Explain the assignment rule in simple words
2.2 For each of the assignment criteria, provide a brief circumstance that likely motivated Novaria's government to conclude this was the best allocation.
2.3 What is the value of N_0?
2.4 What is the value of N_1?
2.5 Create a variable for $Y^{obs}(W)$, and name it `yobs_w`. With this variable:
  (a) Compute the value of $\bar{Y}^{obs}(1)$.
  (b) Compute the value of $\bar{Y}^{obs}(0)$.
2.6 Write down both expressions $\bar{Y}^{obs}(\cdot)$ more formally, in terms of summations.
2.7 Your team knows that $ATE = E(y_i^{obs}|W = 1) - E(y_i^{obs}|W = 0)$ but they don't know why, or how to estimate it from our sample.
  (a) What is $\widehat{ATE}$?
  (b) Justify your answer in terms of a famous mathematical theorem:
    i. $\square \xrightarrow{\square} E(y_i|W = 1)$
    ii. ...
    iii. Therefore the estimate ...
  (c) Compute $\widehat{ATE}$ using R.
  (d) **Reflect on the result**. How does it compare to your answers in 1.6b and 1.6c? At this point, do we have any way to diagnose the accuracy of this result?
2.8 Again, let's imagine we can observe potential outcomes. In the lecture, it was shown that the ATE can be decomposed in *treatment on the treated* and *selection bias*. Write down that expression and estimate the values of:
  (a) treatment of the treated
  (b) selection bias

(c) each individual term in *selection bias*

2.9 What do these values imply for the experiment's design? Is the value for 2.8a a potentially good $\widehat{ATE}$? What would be omitted if we were to only consider this value?

Case study O.7.2

Evaluating AI-assisted learning on student outcomes (continued)

Instructions: Work in pairs or groups of 3; answer the questions as concisely as possible. Type your answers in a single shared document or in the format required by the instructor.

One of you must set up a blank Colab notebook to work on, and share it with the rest. Save any figures or output, and incorporate as required by the instructor.

This is a direct continuation of G.7.3, thus we will work under the context you already have. Continue to use `students` when necessary to answer the questions.

Scenario: You let the Prime Minister know the treatment allocation for the Project Mentor experiment you had previously agreed on is problematic. He hires a team of consultants to help you sort this design problem out, as well as polish other details of the experiment. The following exercises are contain some of the questions asked during the meetings held with the consulting team and the Prime Minister.

Meeting 1

1.1 Firstly, you are asked to explain generally why you cannot get a credible average treatment effect from the current treatment allocation. How would outcomes be different we were to scale up the program nationally? (use your "secret" knowledge of the potential outcomes)

1.2 You are asked to provide an alternative assignment that would create two groups equally representative of 6th-graders nationally. Create such assignment variable under the name `T`, and also create `yobs_t`, the outcome we would observe under this assignment. Calculate the $\widehat{ATE}$. How does this result compare to 2.8a in G.7.3? Without making any further calculations, what do you think the treatment effect among private schoolers will be?

1.3 The Prime Minister doesn't believe that your new assignment created comparable groups. One way to provide evidence of balance, is showing the groups have the same composition of public and private schoolers. Formally show there is no evidence to reject the composition is the same. Be as conservative as possible with the variance.

1.4 Imagine everyone can observe potential outcomes. From the definition of ATE, show treatment and control are comparable more decisively.

Meeting 2

While more convinced of your new assignment, the Pime Minister still insists asking for permission to private schools is impractical and will delay matters. Conveniently, the consulting team asks two questions:

- Whether attending a private/public school in Novartia actually creates systematic differences between students; particularly, differences related to the outcome. This has not been formally shown.
- Whether there may be differences in treatment effects between public and private students (e.g. the mean effect is larger for any), as these differences would justify different rollouts.

To provide evidence in favor or against these questions, bear in mind we have to start from the following assumption:

$$y_i(0)|W = 0 \sim Distr(\mu_0,\ \sigma^2_{0,0})$$
$$y_i(1)|W = 0 \sim Distr(\mu_1,\ \sigma^2_{1,0})$$
$$y_i(0)|W = 1 \sim Distr(\nu_0,\ \sigma^2_{0,1})$$
$$y_i(1)|W = 1 \sim Distr(\nu_1,\ \sigma^2_{1,1})$$

Where $\sigma^2_{0,0} \neq \sigma^2_{0,1} \neq \sigma^2_{1,0} \neq \sigma^2_{1,1}$

2.1 In your own words what do these assumptions mean? What do they entail when it comes to testing hypotheses? According to the lecture, what should we assume about the correlation among these random variables if we want to be conservative?

2.2 Answer the question about systematic differences in outcomes between public and private schools with the evidence you have.
(a) State the appropriate null hypotheses and their alternates.
(b) Test them with the appropriate statistic (estimate any parameters you don't know).
(c) Tie your conclusions directly to the question with 95% confidence.

2.3 Answer the question about differences in treatment effects.
(a) State the appropriate null hypotheses (test equality).
(b) Make inference on $\hat{\nu}_1, \hat{\nu}_0, \hat{\mu}_1, \hat{\mu}_0$ accordingly.
(c) Tie your conclusions directly to the question with 95% confidence.

Meeting 3

Satisfied with your answers, the Prime Minister and consultants approach you with some final questions about the program's broader implications:

3.1 **SUTVA violations** In your own words, briefly explain the Stable Unit Treatment Value Assumption (SUTVA). Could implementing Project Mentor violate this assumption in Novaria? Provide a specific scenario illustrating such a violation clearly. How might these externalities affect the accuracy of your estimates?

3.2 **Alternative policies with proven outcomes (e.g. TaRL)** The consultants suggest evaluating cheaper alternatives, like Teaching at the Right Level — targeting teaching to each student's current skill level without advanced technology; human mentors, complementary material, smaller traditional groups (more teachers), among others. These alternatives currently have more robust evidence of their effects.

(a) What, if any, are some differences between the theory of change underlying TaRL and that of Project Mentor?

(b) If differences exist, briefly discuss how you would test them within an experimental design, clearly describing treatments, assignment and measured outcome(s).

(c) Apart from the treatment effects, what else is necessary if we wanted to fairly compare any known TaRL intervention to Project Mentor in terms of efficiency?

(d) In this comparison, how important do you think scale would be? Briefly describe how costs for one and the other would behave.

3.3 **Non-compliance** Not every school or student may strictly follow the treatment assignment. Describe one realistic scenario of non-compliance in this project. Explain briefly how such non-compliance might bias the estimated treatment effect. Suggest one practical strategy to reduce or mitigate non-compliance.

14.310x Flipped Classroom
Week 8 Instructions

Some title

Checklist

- ☐ Watch the linear models with R: `lm` tutorial[4] (Requirement)
- ☐ Complete Guided Case Study 4 (Session 1)
  - ☐ Get the `bike_rentals` dataset here
- ☐ Complete Guided Case Study 5 (Session 2)
  - ☐ Get the `student_performance` dataset here
- ☐ Complete Guided Case Study 6 (Session 2)
  - ☐ Get the `kc_house_data` dataset here
- ☐ (Optional) Explore **Review Notes** and further **Readings** available.

---

[4] **Module 8** > **Introduction to the Class lm** in the online component of the course.

Guided case study G.8.4

Part A - Bike rentals

Instructions: Work in pairs or groups of three. Solve the exercises for Part A of *Linear Regression* in Colab. Work collaboratively in a single Notebook.

Notebook: Part A

Dataset: bike_rentals.csv

Type your answers in the notebook, and submit your work per the instructor's requirements.

Guided case study G.8.5

Part B - Student performance

Instructions: Work in pairs or groups of three. Solve the exercises for Part B of *Linear Regression* in Colab. Work collaboratively in a single Notebook.

Notebook: Part B

Dataset: student_performance.csv

Type your answers in the notebook, and submit your work per the instructor's requirements.

Guided case study G.8.6

Part C - KC Housing Data

Instructions: Work in pairs or groups of three. Solve the exercises for Part C of *Linear Regression* in Colab. Work collaboratively in a single Notebook.

Notebook: Part C

Dataset: kc_house_data.csv

Type your answers in the notebook, and submit your work per the instructor's requirements.

Additional resources

- Drive: Review Notes
- Drive: Readings

14.310x Flipped Classroom

Week 9 Instructions

Some title

Checklist

- ☐ Complete Coding Lab 5 (Session 1)
- ☐ Complete Guided Case 7 (Session 2)
- ☐ Review the reference material as needed (Optional)

Coding Lab C.9.5

Coding Lab 5— Further regression tools

Instructions: Work individually. Complete all the coding lab's exercises.

Notebook: Further regression tools

Upon completion, submit your work in the format required by the instructor.

Guided Case Study G.9.7

Regression Discontinuity: Replicating *Going to a better school* (Pop-Eleches & Urquiola, 2013)

Instructions: Work in pairs or teams of three. We will go over a section of *Going to a better school: Effects and Behavioral Responses* to examine outcomes differences of children attending higher achievement schools.

Notebook: Guided Case 7— RDD paper replication

Dataset: Schools.dta

Paper: (Pop-Eleches & Urquiola, 2013)

Scenario: In 2002, the Romanian Ministry of Education centralized the high school admissions system, ranking students by their standardized test scores and matching them to schools in descending order of achievement. This reform generated a compelling natural experiment: some students who barely made it into higher-achieving schools were nearly identical, academically, to those who just missed the cutoff. In this guided case study, you will analyze data from this setting to investigate whether attending a better school causally affects student outcomes. You'll explore the design of the regression discontinuity strategy used by Pop-Eleches and Urquiola (2013), and replicate parts of their analysis using real admissions and outcomes data from Romanian high schools.

A Colab notebook has been set up with the full case directions therein. Submit your coursework in the format required by your instructor.

Additional materials and references

References:

Drive:

- (Pop-Eleches & Urquiola, 2013)
- Causal inference textbook
- Econometrics textbook

Additional materials:

Drive:

- Review notes week 9

Bibliography

Pop-Eleches, C., & Urquiola, M. (2013). Going to a better school: Effects and behavioral responses. *American Economic Review*, *103*(4), 1289–1324. https://doi.org/10.1257/aer.103.4.1289

14.310x Flipped Classroom
Week 10 Instructions

Some title
